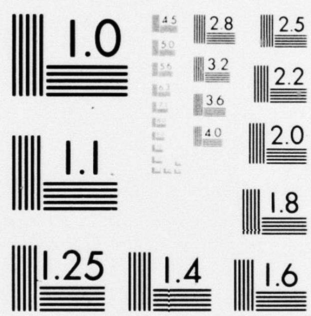
AD-A042 888 TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING F/G 12/1
ASYMPTOTIC PROPERTIES OF CLUSTERING ALGORITHMS,(U)
JUN 77 L P DEVROYE, T J WAGNER AF-AFOSR-2371-72
UNCLASSIFIED AFOSR-TR-77-0905 NL

1 OF 1
ADA042888

END
DATE
FILMED
9-77
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Reprinted from the Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Troy, NY, June 6-8, 1977.

ASYMPTOTIC PROPERTIES OF CLUSTERING ALGORITHMS

L.P. Devroye and T.J. Wagner*
Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712

Abstract

Suppose a sample of size n is observed from the d-dimensional density f. Conditions are given which insure that a single-linkage clustering algorithm can asymptotically find the decomposition of the support of f into connected closed sets.

Clustering is the process of grouping similar objects. For our purposes the objects to be grouped can be thought of as a set of d-dimensional vectors and a clustering algorithm can be thought of as any scheme for partitioning this set into subsets called clusters. Our paper analyzes the asymptotic performance of clustering algorithms for a simple probabilistic model with the result that versions of a single-linkage clustering algorithm are shown to be asymptotically effective. Excellent summaries of previous work in clustering are contained in Hartigan and Dorofeyuk[1,2], while a more technical and thorough description of our results may be found in Devroye and Wagner[3].

Let X be a random vector with values in $\mathbb{R}^d$ and a probability density

$$f = \sum_1^M \pi_i f_i \tag{1}$$

where $\pi_i > 0$, $1 \le i \le M$, $\sum_1^M \pi_i = 1$ and $f_1,\dots,f_M$ are probability densities. If f_i has support C_i, $1 \le i \le M$, then we assume that

(a) C_i is connected, $1 \le i \le M$,
(b) $C_1,\dots,C_M$ are disjoint, and (2)
(c) C_i is bounded, $1 \le i \le M$.

The supports $C_1,\dots,C_M$ may be thought of as the clusters chosen by nature. In particular, if independent observations are made on (1) then $C_1,\dots,C_M$ determine a natural partition of these observations. However, suppose that the statistician assumes only that (1) and (2) hold for some M, $\pi_1,\dots,\pi_M$ and $f_1,\dots,f_M$, and, in place of specific knowledge of f, has a sample size n from (1), say $X_1,\dots,X_n$. The question that concerns us here is how the statistician can asymptotically obtain the same grouping of observations on (1) as he would if he knew $C_1,\dots,C_M$.

From the sample $X_1,\dots,X_n$, the statistician will, for his clustering algorithm, construct a partition $A_1,\dots,A_L$ of $\mathbb{R}^d$. Future observations, that is, observations from (1) which are independent of those in his sample, will then be grouped together if they fall in the same set A_ℓ. For this reason, we shall also refer to the sets $A_1,\dots,A_L$ as clusters. In the vast clustering literature, concentration is focused on grouping the sample $X_1,\dots,X_n$ and the sets of the partition of $X_1,\dots,X_n$ determined by $A_1,\dots,A_L$ are usually referred to as clusters. Concentrating on partitioning $X_1,\dots,X_n$ seems warranted, for example, in clustering problems arising in paleontology studies where new observations are not expected. However, in medical situations, such as trying to cluster the types of shock for emergency care purposes, the statistician is interested in the performance of the algorithm on future observations. Our model is directed toward this type of situation.

Referring to Figure 1 there are three natural clusters but the algorithm with the sample $X_1,\dots,X_n$ has yielded four clusters in (a), three in (b) and two in (c). How does one measure the performance of the algorithm on future observations? Agreeing that what we call each cluster C_i is unimportant as long as we give one unique label to each C_i we see that the probability of misclassification becomes

$$L_n = \min_g \sum_{i=1}^M \pi_i \int_{A^c_{g(i)}} f_i(x)dx \tag{3}$$

where the minimum is taken over all one-to-one functions g: $\{1,\dots,M\} \to \{1,\dots,\mathrm{Max}(M,L)\}$ and, if $M > L$, we put $A_{L+1} = \dots A_M = \phi$. In particular, if C_i is contained in some A_j and each A_j contains at most one C_i then $L_n = 0$. It should be stressed that L_n is a random variable which depends on $X_1,\dots,X_n$ and whose value is just the frequency of observations misclassified when a large number of new observations are classified with the partition $A_1,\dots,A_L$.

Our interest here is finding what properties are necessary for clustering algorithms to insure that $L_n \to 0$ with probability one. The following clustering algorithm, a version of the familiar single-linkage algorithms, has this property with some slight additional assumptions on f. More extensive results for other algorithms and assumptions may be found in Devroye and Wagner[3].

If $r > 0$ connect the two points X_i, X_j if $d(X_i,X_j) < r$, $1 \le i,j \le n$. Call two points X_k, X_ℓ connected if there exists a sequence $Y_0,\dots,Y_m$ from $\{X_1,\dots,X_n\}$ with $Y_0 = X_k$, $Y_m = X_\ell$, and Y_{i-1}, Y_i, connected, $1 \le i \le m$. The set $\{X_1,\dots,X_n\}$ is then partitioned into connected subsets $K_1,\dots,K_L$. A partition $A_1,\dots,A_L$ of $\mathbb{R}^d$ is obtained from $K_1,\dots,K_L$ by putting the point $x \in \mathbb{R}^d$ into A_j if the closest point to x from $X_1,\dots,X_n$ is in K_j (ties are broken

*Supported in part by AFOSR Grant 72-2371.

(See 1473)

arbitrarily).

Theorem. If $r = r_n$ satisfies

$$\begin{aligned} &\text{(i)} \quad nr_n^d/\log n \to \infty \\ &\text{(ii)} \quad r_n \to 0 \end{aligned} \tag{4}$$

and if, for some a,b > 0,

$$\inf_{x \in \bigcup_1^M C_i} \int_{S(x,\rho)} f(x)dx \geq a\rho^d, \; 0 \leq \rho \leq b, \tag{5}$$

where $S(x,\rho)$ is the sphere centered at x with radius ρ, then

$$L_n \to 0 \quad \text{w.p.1.}$$

Proof. We recall that the support C of a density f is the smallest closed set with the property that $\int_C f(x)dx = 1$. In particular, the C_i are closed sets. Because the C_i are bounded,

$$\inf_{x \in C_i, y \in C_j} \|x-y\| \geq \delta > 0$$

whenever $i \neq j$. We assume that n is so large that $r_n < \delta$ (use (4(ii))). Suppose that $X_1,\ldots,X_n$ is such that every sphere $S(x,r_n/3)$ contains at least one of the X_i for $x \in C = \bigcup_1^M C_i$.

If we can show that

(i) whenever $C_i \cap A_j \neq \phi$ and $X_\ell \in C_i$, then $X_\ell \in A_j$, and

(ii) whenever $C_i \cap A_j \neq \phi$ and $X_\ell \notin C_i$, then $X_\ell \notin A_j$,

then we know that M = L and $\bigcup_{i=1}^M C_i = \bigcup_{i=1}^M (C_i \cap A_{g(i)})$, for some one-to-one mapping g: $\{1,\ldots,M\} \to \{1,\ldots,M\}$, which in turn implies that

$$0 \leq L_n \leq \sum_{i=1}^M \pi_i \int_{A^c_{g(i)}} f_i(x)dx = 0$$

and

$$P\{L_n > 0\} \leq P\{\inf_{x \in C} \mu_n(S(x,r_n/3)) = 0\} \tag{6}$$

where μ_n is the empirical measure for $X_1,\ldots,X_n$.

Let us now prove (i) and (ii). Property (ii) is trivial since $r_n < \delta$ and $X_j \in \bigcup_1^M C_i$ for all j with probability one. For property (i), we need only show that for any x in C_i, and any $X_j \in C_i$, there exists a sequence $Y_1,\ldots,Y_\ell$ from $X_1,\ldots,X_n$ with $Y_1 = X^{(1)}$, $Y_\ell = X_j$, $\|Y_{k+1} - Y_k\| \leq r_n$, $1 \leq k < \ell$, where $X^{(1)}$ is the nearest neighbor to x among $X_1,\ldots,X_n$.

Since $S(x,r_n/3)$ contains one X_k, and since $r_n < \delta$, we know that $X^{(1)}$ belongs to C_i as well, no matter what x is picked in C_i. By the connectedness of C_i, we can find $\{x_1,\ldots,x_\ell\} \subseteq C_i$ with $x_1 = X^{(1)}$, $x_\ell = X_j$, and $\|x_{k+1} - x_k\| < r_n/3$, $1 \leq k < \ell$. Thus, since every $S(x_i,r_n/3)$ contains one of the X_k's, we know that there are $Y_k \in S(x_k,r_n/3)$, $1 \leq k \leq \ell$, $Y_1 = X^{(1)}$, $Y_\ell = X_j$. Also, $\|Y_{k+1} - Y_k\| \leq \|Y_{k+1} - x_{k+1}\| + \|x_{k+1} - x_k\| + \|Y_k - x_k\| < r_n$. This concludes the proof of (i).

As for (6), because the C_i are bounded, we can find a grid $\{y_1,\ldots,y_N\} \subseteq C$ with the property that for every $x \in C$ there exists an y_i with $\|y_i - x\| \leq r_n/12$. Such a grid contains at most γ/r_n^d points where $\gamma > 0$ is a constant depending upon d, $\|\cdot\|$, and the diameter of C. If $r_n/12 < b$, then

$$\inf_i \int_{S(y_i,r_n/6)} f(z)dz \geq \inf_{x \in C} \int_{S(x,r_n/12)} f(z)dz \geq \alpha\left(\frac{r_n}{12}\right)^d .$$

Also, if $\inf_{x \in C} \mu_n(S(x,r_n/3)) = 0$, then $\mu_n(S(y_i,r_n/6)) = 0$ for all i so that

$$\begin{aligned} P\{L_n > 0\} &\leq P\{\inf_{x \in C} \mu_n(S(x,r_n/3)) = 0\} \\ &\leq \sum_{i=1}^N P\{\mu_n(S(y_i,r_n/6)) = 0\} \\ &\leq \left(\gamma/r_n^d\right)\left(1 - \inf_{x \in \bigcup_1^M C_i} \int_{S(x,r_n/12)} f(z)dz\right)^n \\ &\leq \left(\gamma/r_n^d\right)\left(1 - r_n^d\,\alpha/12^d\right)^n \\ &\leq \frac{\gamma}{r_n^d} e^{-\alpha n r_n^d/12^d} . \end{aligned}$$

By the Borel-Cantelli lemma and (4)(i), we have that $\sum_n P\{L_n > 0\} < \infty$, completing the proof of the Theorem.

Q.E.D.

References

1 J.A. Hartigan, Clustering Algorithms, John Wiley, N.Y., 1975.

2 A.A. Dorofeyuk, "Automatic classification algorithms (review)," Automation and Remote Control, 32, pp. 1928-1958, 1971.

3 L.P. Devroye and T.J. Wagner, "Asymptotic Properties of Hierarchical Clustering Algorithms," Technical Report, January 1977, Dept. of Electrical Engineering, University of Texas, Austin, TX.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR-77-0905	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ASYMPTOTIC PROPERTIES OF CLUSTERING ALGORITHMS		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L.P. Devroye and T.J. Wagner		8. CONTRACT OR GRANT NUMBER(s) AFOSR 72-2371
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Texas Department of Electrical Engineering Austin, TX 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332		12. REPORT DATE June 1977 13. NUMBER OF PAGES 2
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON PATTERN RECOGNITION AND IMAGE PROCESSING, Troy, NY, pp. 321-322, June 6-8, 1977.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Clustering; Density Estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Suppose a sample of size n is observed from the d-dimensional density f. Conditions are given which insure that a single-linkage clustering algorithm can asymptotically find the decomposition of the support of f into connected closed sets.

401997 (over)